RAND

Rewarding Schools Based on Gains: It's All in How You Calculate the Index and Set the Target

Brian Stecher and Jeremy Arkes

DRU-2532

April 2001

RAND Education

Contents

Tables

Figures

Introduction[1]

President Bush has proposed that all states implement a test-based accountability system like the Academic Excellence Indicator System (AEIS) that exists in his home state of Texas. In that system, and in similar systems in California, Florida, Kentucky, North Carolina, and other states, each school receives a rating based on its students' test scores. Schools with high ratings receive formal recognition and (often) financial rewards. Schools with consistently low ratings must engage in focused improvement efforts which may ultimately lead to reconstitution. In some states, the parents of students from schools with poor ratings are allowed to transfer their children to other schools. The key ingredients in such test-based accountability systems are an index, a target, and a series of consequences. The index is the scale used to rate the school's performance for the purpose of accountability. It is derived from the performance of students on tests as well as other factors. The target is the index value or values used to determine a school's status in the accountability system. The target may be defined in terms of the absolute value of the index, growth in the index, or a combination of the two. The school's standing with respect to the target leads to positive or negative consequences.

The purpose of this paper is to present some empirical evidence about how the allocation of rewards depends on the methods used to construct the accountability formulas. As far as we know, there is little published research on this topic, and state policymakers have limited information on which to base choices. In particular, we examine the effect on school rewards of changes in three features of California's accountability system: the way student test scores are combined into a school index, the way targets are set, and whether the system has additional targets for particular subgroups of students. In the first case, we compare an index defined in terms of a linear metric (mean National Curve Equivalent gains) to one that is more "progressive" (i.e., one that allocates more points for gains among low-scoring students than for gains among high-scoring students. In the second case, we compare three methods for defining gain targets—a fixed amount of improvement, a fixed percentage improvement, and a "percentage of distance to target." Finally, we examine the effects of imposing similar conditions on subgroups of students as well as the school as a whole. We use 1998-99 and 1999-2000 data from California and from the Los Angeles Unified School District as the basis for our analyses. This study suggests that small differences in computational formulas can have large effects in terms of consequences for schools.

Background

Accountability Systems

Many states have implemented test-based accountability systems (TBA) in the past few years and, although they differ in a number of details, they have broad similarities. The Texas system is typical of current TBA. Under the Academic Excellence Indicator System (AEIS) that exists in Texas, each school receives a rating (low-performing, acceptable, recognized, or exemplary) based on the performance of its students on the Texas Assessment of Academic Skills (TAAS), and the school's attendance and dropout rates.[2] The TAAS tests cover reading, mathematics, and

---

[1] This paper was presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA, April 13, 2001.

[2] There is also a category for special circumstances under which a school may not be rated or may be rated under alternative criteria.

writing, and they are administered to students in grades three through eight and grade ten.[3] For a school to reach the acceptable level at least 50 percent of students must pass the TAAS test in each subject, the school's dropout rate must be 6 percent or less, and attendance for the year must be at least 94 percent. The dropout and TAAS requirements apply to all students and to subgroups of students based on race/ethnicity and economic status. The state provides small financial rewards for schools whose ratings are high.[4] In 1999 qualifying schools received from $500 to $5,000 based on an allocation of $4.58 per student. In contrast, schools that continue to perform poorly are subject to increasing sanctions, beginning with site visits from a peer review team and the development of a school improvement plan. Since 1995, the state has provided funds to allow parents with children in poor-performing schools to transfer their students to higher-performing schools (even schools outside the district boundaries).

Other states have adopted accountability systems that are broadly similar to Texas, although each of these systems differs from Texas in interesting ways. For example, the Kentucky system is based on two-year rather than annual reward cycles. California schools must show 5 percent improvement annually to be eligible for rewards. In Florida, the state has set aside funds to provide vouchers to parents of students in schools that fail to meet the requirement two years in a row. In North Carolina, targets are set on the basis of the average gain of all students in the state, and then adjusted for the initial scores of students in each school. In Washington, the Commission charged with developing a model for accountability recommended that special emphasis be placed on a school's success in improving the performance of students scoring at the lowest level. In Massachusetts, schools are judged both in terms of absolute performance and in terms of improvement, and these judgements are made on the basis of schools' average performance over two years. These examples illustrate the flexibility states have in defining specific accountability mechanisms.

At their core each of these accountability systems has three parts—an index, a target, and a series of consequences. The index is the scale used to summarize a school's performance for the purpose of accountability. It can be nominal (e.g., meets or does not meet standard), ordinal (e.g., a grade of A to F) or interval (e.g., percent of students exceeding the mean national percentile rank on a test). The scale can be derived from a single measure (e.g., a standardized achievement test) or a combination of measures (e.g., a weighted sum of test scores, attendance, and dropout rates). It can be defined in the same metric as the underlying measure (e.g., national percentile rank) or it can transform that scale in some manner (e.g., assign students to levels and award points based on level). It can be based on the performance of all students or a subset of students (e.g., only students in a designated grade level).

The target is the index value or values used to determine a school's status in the accountability system. The target may be defined in terms of status or growth or a combination of the two. For example, a school might have to reach a specific level of performance (e.g., a fixed percent of students meeting the criteria for "proficient"). A school might have to show a fixed amount of improvement (e.g., an increase in the index of 5 percent over the previous year). Or a school's target might be a function of status and growth (e.g., an increase of 5 percent of the difference between the current value and a long-range goal). Gains can be computed as the difference between the school's indices in two successive years ignoring differences between the students from one year to the next ("cross-sectional") or as the difference between the indices of matched students who attended during both years ("longitudinal").

---

[3] Writing is only tested in grades four, eight and ten. Tenth grade students can also qualify by passing three end-of-course examinations—English II, algebra I and either U.S. history or biology.

[4] In 1999 the Texas legislature appropriated $5 million for rewards for schools rated highly that had demonstrated significant gains in student test scores. In addition, the highest-performing schools and districts are exempted from certain regulations.

The school's standing with respect to the target leads to positive or negative consequences. Schools whose indices surpass the target receive public recognition as a successful school, and they may also receive tangible rewards, including relative large amount of money. There may be one level of reward or multiple levels. Schools whose indices fail to achieve the target are labeled as deficient, and other sanctions may be imposed. In many states the sanctions escalate if the school continues to score poorly year after year. Sanctions often begin with some type of focused school improvement effort (forced review by an external evaluator, assignment of a designated educator to help with planning, etc.). Repeated failure may lead to stiffer sanctions, such as loss of funds, dismissal of administrators or staff, provision of vouchers to students, reconstitution, etc.

These three features are the core of any school-level accountability system, but they are not the only elements that may be included. Some systems impose additional conditions to ensure that subgroups are making equivalent progress. For example, in California there is a target for the school as a whole, and there are separate targets for each significant subgroup of students (80 percent of the school's target). To be eligible for rewards schools must meet their overall target and their subgroup targets. It is also possible that an accountability system might include adjustments to either the index or the target to reflect the demographic characteristics of the students served by the school. For example, a school might be given a predicted score based on student characteristics and receive a reward if they exceed their predicted scores.

The California Accountability Program

Our analyses are conducted on data from California, but TBA systems in many states are similar enough to make the results of more general value. In 1999, the California legislature created a formal educational accountability system for the state. The act set out parameters for a school-based accountability system that included an Academic Performance Index (API), a remedial program for under-performing schools, and a program for improving schools. The law required that the index consist of test scores from the Student Testing and Reporting (STAR) program, pupil and certified personnel attendance rates (if accurate data were available), graduation rates (if accurate data were available), and other statewide test results (when available, valid and reliable). Pupil test results had to constitute at least 60 percent of the index. The law required that the state rank schools by API value, by growth in API, and by growth in API compared to similar schools. Furthermore, the Superintendent of Public Instruction was directed to establish growth targets that should be "a minimum of 5 percent annually." The law permitted the targets to be greater than the minimum for the lowest-performing schools. Schools are also required to show comparable improvement for numerically significant subgroups of students. Subgroups are defined in terms of racial/ethnic group and socioeconomic status.[5]

The law also includes rewards and sanctions based on API gains. Both the reward program and the sanctions are still being developed, and are likely to evolve over time. Initially, the sanctions will begin with an intervention program for schools making inadequate API gains. The Immediate Intervention/Under-performing Schools Program (II/USP) involves structured planning for improvement under the guidance of an experienced external evaluator. Schools receive $50,000 to undertake this process, and then they have a year to show improvement. Schools that fail to meet their targets after a year are subject to a range of actions, including reassignment of school personnel, negotiation of site specific amendments to teacher bargaining unit contracts, and other interventions. Schools that fail to meet their targets within 24 months and fail to demonstrate "significant growth" will be the object of direct intervention by the state Superintendent, which can lead ultimately to closing the school.

---

[5]Alternative accountability systems are to be established for schools with fewer than 100 pupils, alternative schools, county schools and other a few other types of non-traditional schools.

High-achieving schools are eligible for the Governor's Performance Award program, which can include monetary and non-monetary awards. In 2000, legislators created two additional reward programs. All together these programs set aside over $750 million for rewards to school based on their API growth from 1999 to 2000. California schools that surpass their school-wide target and meet all subgroup targets receive $63 per student; they also receive an additional sum based on the number of staff. Schools that surpass twice their target are eligible for additional teacher bonuses of up to $25,000 per teacher.

Computing the API

The API is based entirely on students' scores on the Stanford-9 test, which is administered to all students in grades two through eleven each spring. Elementary and middle school students are tested in reading, language, spelling and mathematics. High school students are tested in reading, language, mathematics, science and social studies. Students are classified into five performance bands based on their national percentile rank on each subject matter tests of the Stanford-9.[6]

The API for a school is calculated in three steps. First, determine what percent of students scored in each quintile of the national distribution for each subject area tested. Second, multiply those percentages by the weighting factors shown in Table 1 for each subject area and sum them up. Third, calculate a weighted average of the scores for the subject areas using the weights shown in Table 2. The choice of API scale (200 to 1000) was arbitrary, but the weighting factors were chosen purposefully to reward improvement over time at the bottom of the distribution more than improvement at the top. Raising a student from the 15th NPR one year to the 25th NPR the next increases that student's contribution to the API by 300, while raising a student from the 75th NPR one year to the 85th NPR the next results in a gain of only 125 points.

Table 1
Converting Stanford-9 Scores to API scores

Performance Band	Weighting Factors
80-99th NPR	1000
60-79th NPR	875
40-59th NPR	700
20-39th NPR	500
1-19th NPR	200

Table 2
Subject Matter Weights

Subject	Grades 2-8	Grades 9-11
Reading	.30	.20
Language	.15	.20
Spelling	.15	--
Mathematics	.40	.20
Science	--	.20
Social Studies	--	.20

[6] The scale score values that define the boundaries between bands are fixed at the 1999 levels; they will not change in future years if the test norms change.

Qualifying for Rewards

The law established 5 percent as the minimum threshold for awards, but it did not define exactly how 5 percent was to be interpreted. The Board of Education reviewed a number of different operational definitions and adopted a "distance to target" approach. First, they set an interim performance target of 800 points.[7] To qualify for rewards from one year to the next, a school's API must improve by 5 percent of the distance between the first year's score and 800. The target for schools whose first year score was between 780 and 800 was set at one point. Schools scoring at 800 or above also had to improve by one point to be eligible for rewards.[8]

The qualification formula was selected to achieve three main goals. First, the Board of Education wanted to provide clear incentives for schools to improve. This formula required all but the very highest-performing schools to show some growth. Second, the Board wanted to set the bar higher for schools that were doing poorly. It was very important that the accountability system create pressure to narrow the gap between schools whose students were doing well and schools whose students were doing poorly. The "distance to target" approach created greater challenges at the bottom of the distribution than at the top. Third, the state wanted all schools to have reasonable chances of earning rewards. They did not want a substantially greater percentage of low-achieving schools or high-achieving schools to qualify for rewards. In 2000, 78 percent of elementary schools qualified for rewards, and the reward winners were distributed roughly equally across all ten deciles of base year performance. (Approximately 60 percent of middle schools and 40 percent of high schools qualified for rewards.)

Methods

Data Sources

We obtained school-level information on all public elementary, middle, and high schools from the California Department of Education. The information includes the demographic characteristics of the students and the API score data for 1999 and 2000. The demographic information includes the number and percent of tested students who are in seven different racial/ethnic groups. It also includes the percent of tested students who are participating in the free or reduced-price lunch program, who are English language learners, and who are attending the school for the first time in the given year. The file also includes information on parents' education, but these variables have missing values for large percentages of students in certain schools. Finally, there are detailed data on API scores and targets for the schools and for each significant racial/ethnic subgroup in the school. Table A-1 in the Appendix summarizes some of the demographic information on the schools in our sample.

We also obtained information on school performance in 1999 and 2000 from the Los Angeles Unified School District. The LAUSD data are helpful because gains are computed in terms of a more uniform scale, the change in Normal Curve Equivalent scores. For each school, the file contained the mean NCE change from 1999 to 2000 of all students tested in 2000 who were enrolled anywhere in the district in 1999. About 80 percent of the students included in the API computations produced by the state are included in this matched student file.

---

[7] David Rogosa of Stanford University, examined the distribution of API scores in 1999 and 2000 and found that an API of 800 was roughly equivalent to having 72.5 percent of the students in the school scoring above the national median. If exactly one-half of the students in a school scored above the 50th NPR the API would be roughly 660.

[8] In addition, at least 95 percent of students must take the Stanford-9 test. There are other technical criteria having to do with parent waivers, changes in the population of the school, irregularities with test administration, etc. that we do not consider in this analysis.

California API Performance

Table 3 summarizes the API scores of California elementary, middle and high schools for 1999 and 2000. It shows the statewide mean scores and targets and the percentage of schools that met their school-wide targets, their subgroup targets, and both sets of targets. On average, the three levels of schools had about the same base year scores and the same targets, but the elementary schools showed much greater improvement from 1999 to 2000. Thus, they were more likely to meet the school-wide targets and subgroup targets. Overall, 78.5 percent of elementary schools met all targets and were eligible for rewards, while only 59.8 and 42.9 percent of the middle and high schools, respectively, were eligible for the rewards.

Table 3
API Performance of Schools in California by Level

	Elementary schools N=4609		Middle schools N=1066		High schools N=672	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
1999 API score	633.6	136.8	636.3	124.3	635.6	103.3
2000 API score	672.5	130.6	658.1	122.8	649.3	101.1
Growth target	8.8	6.1	8.6	5.7	8.5	4.8
School-wide target met	88.1%		73.6%		58.2%	
Subgroup targets met	80.5%		63.1%		45.7%	
Met all targets	78.5%		59.8%		42.9%	

Distribution of Rewards by Base Year Achievement Level and School Lunch Program Participation

In Table 4, we show how scores and targets vary across schools based on 1999 API score. Table 5 provides a similar comparison of school performance based on family income—in this case schools are divided into quartiles based on the percent of students who are on the free or reduced-price lunch program. Comparing scores, targets, and rewards this way allows us to examine the allocation of rewards across schools of different achievement and family income levels.

It is evident in Table 4 that as the base year API increases the improvement in API scores and the percent improvement in scores decreases. This is true for all levels of schooling: elementary, middle and high schools. Despite their larger API gains, the lower-scoring schools are not more likely to meet their targets because their targets are much greater. Thus, the percent of schools that met all targets is fairly similar across all deciles. Specifically, for the lowest eight deciles, the percent of schools that met all targets stays in a narrow range of 75.7 to 78.7 percent. For the top two deciles, the percent increases to 83.2 and 87.4 percent, respectively. This pattern is also evident for middle schools, but not for high schools. The cause of the increase in the higher two deciles appears to be greater success in meeting subgroup targets, which is directly related to the fact that schools in the upper two deciles have fewer subgroups of significant size. A similar pattern is true when schools are divided in terms of student participation in the school lunch program (see Table 5).

Table 4
Success at Meeting API Growth Targets by 1999 API Score Decile by Level

	Deciles based on 1999 API score									
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Elementary Schools										
1999 API	413.2	475.9	523.0	567.7	610.5	652.7	695.5	742.4	791.6	862.6
Growth target	19.4	16.2	13.9	11.7	9.4	7.4	5.2	2.9	1.1	1.0
Improvement from 1999-2000	48.1	46.2	48.8	46.4	45.9	39.4	35.3	32.8	28.1	18.0
Percent improvement	11.7	9.7	9.3	8.2	7.5	6.1	5.1	4.4	3.6	2.1
Whether school-wide target met	83.1%	83.9%	87.2%	87.3%	90.2%	90.1%	91.0%	89.9%	89.9%	88.7%
Whether subgroup targets met	79.4%	78.1%	77.0%	76.6%	79.6%	76.7%	76.8%	78.9%	86.0%	96.3%
Whether all targets met	77.4%	76.2%	76.3%	75.7%	78.7%	75.8%	75.9%	78.5%	83.2%	87.4%
Middle Schools										
1999 API	428.5	495.9	542.0	581.5	619.2	655.4	690.5	729.5	776.2	847.9
Growth target	18.6	15.2	12.9	10.9	9.1	7.3	5.5	3.5	1.4	1.0
Improvement from 1999-2000	25.2	23.8	24.0	25.2	24.8	22.9	20.7	21.2	18.2	12.5
Percent improvement	6.0	4.8	4.4	4.3	4.0	3.5	3.0	2.9	2.4	1.5
Whether school-wide target met	61.3%	59.6%	67.9%	77.4%	75.7%	81.3%	72.4%	78.0%	83.0%	80.0%
Whether subgroup targets met	58.5%	47.7%	50.0%	62.3%	53.3%	58.9%	63.8%	71.6%	73.6%	92.4%
Whether all targets met	55.7%	44.0%	49.1%	61.3%	53.3%	58.9%	61.0%	68.8%	70.8%	76.2%
High Schools										
1999 API	462.0	518.4	562.2	594.0	620.0	648.8	679.3	708.0	747.4	817.5
Growth target	17.0	14.1	11.9	10.3	9.1	7.6	6.1	4.6	2.7	1.1
Improvement from 1999-2000	21.3	13.4	27.4	13.3	12.3	13.8	15.9	6.1	8.0	6.4
Percent improvement	4.7	2.6	4.8	2.3	1.9	2.1	2.3	0.9	1.1	0.8
Whether school-wide target met	52.2%	41.8%	75.8%	56.5%	53.7%	58.6%	65.6%	58.0%	55.2%	65.2%
Whether subgroup targets met	50.7%	32.8%	45.5%	39.1%	41.8%	40.0%	54.7%	36.2%	43.3%	74.2%
Whether all targets met	47.8%	32.8%	45.5%	36.2%	41.8%	38.6%	53.1%	36.2%	41.8%	56.1%

Table 5
Success at Meeting API Growth Targets by Free/Reduced-Price Lunch Status by Level

	Quartiles based on Free/ Reduced-Price Lunch			
	4th	3rd	2nd	1st
Elementary Schools				
1999 API	479.6	577.5	681.8	797.4
Growth target	16.0	11.2	6.0	2.0
Improvement from 1999-2000	43.4	46.2	36.9	29.1
Percent improvement	9.4	8.2	5.6	3.8
Whether school-wide target met	82.5%	89.0%	89.0%	92.2%
Whether subgroup targets met	75.4%	78.9%	76.4%	91.7%
Whether all targets met	74.1%	77.6%	74.9%	87.6%
Middle Schools				
1999 API	494.8	589.5	679.1	781.4
Growth target	15.3	10.6	6.1	2.3
Improvement from 1999-2000	23.0	22.6	22.5	19.3
Percent improvement	4.8	3.9	3.4	2.6
Whether school-wide target met	62.6%	69.8%	79.8%	82.3%
Whether subgroup targets met	50.7%	55.0%	63.6%	83.1%
Whether all targets met	48.1%	53.9%	62.5%	74.8%
High Schools				
1999 API	526.8	601.6	670.8	741.7
Growth target	13.7	10.0	6.6	3.6
Improvement from 1999-2000	17.7	14.8	13.3	9.4
Percent improvement	3.5	2.5	2.1	1.4
Whether school-wide target met	52.1%	61.4%	58.0%	61.0%
Whether subgroup targets met	42.5%	41.5%	40.7%	57.6%
Whether all targets met	41.3%	40.4%	39.5%	50.0%

Analyses

We are investigating the effect on rewards of changes in the computation of the accountability index and the setting of reward targets. First, we compare an index defined in terms of a linear metric to the current method that allocates more points for gains among low-scoring students. Second, we compare how different targets affect which schools meet the criteria. The options we consider include a fixed amount of improvement, a fixed percentage improvement, and a percentage of the distance to a set target. Third, we examine the effects of changing the subgroup targets. In particular, we are interested in whether the various schemes would be more likely to give rewards to lower- or higher-performing schools (based on 1999 scores) or to schools with more or less affluent students. To examine the effects of these changes, we duplicated the computations used in the current system then applied new targeting schemes to determine how these rules would affect the allocation of rewards.

Changing to a Linear Metric

The first part of our analysis examined the effect on the allocation of rewards of weighting student gains more equally in computing the school index. This analysis shows the effect of the

state's decision to assign more points to gains in the lower end of the distribution than to gains in the higher end of the distribution. There are a number of ways to re-scale the index that would make it more linear with respect to initial achievement. One would be to use the API formula but make the weights the same for all five NPR quintiles. An alternative would be to compute gains at the individual student level using Normal Curve Equivalent (NCE) scores, which is the appropriate scale to use for this purpose.

To see how changing the metric affected rewards, we merged the mean NCE gain data from LAUSD with the API gain data from the California Department of Education. To make the comparisons equivalent we established a new benchmark, the gain in each metric at which one-half of the schools in LAUSD receive awards. These values were 2.43 NCE points and 43 API points. We repeated the process using as a benchmark the gain in NCE scores and in API scores that would result in 70 percent of schools achieving rewards. These targets were 1.60 NCE points and 31 API points. (The LAUSD data do not include information on subgroup scores, so we did not include subgroups in this analysis.) We compare the percentage of schools that met each target based on the school's 1999 California API decile. Because there are fewer schools in the higher deciles in LAUSD, we consolidated the schools into quintiles. From the lowest to the highest quintile, there are 213, 79, 57, 34, and 28 schools in this analysis.

Changing School-wide Targets

Before the current API targets were adopted California considered a number of possible models. In this analysis we examined five different models, which included some of the same options considered by the state. The five models are as follows:

1. The current system: *5 percent of the distance from 1999 API score to 800* (all schools with an 1999 API score above 779 have a target of 1).
2. A target equal to 5 percent of the distance from 1999 API score to 1000 (the maximum score).
3. A target equal to 5 percent of 1999 API score.
4. A fixed target equal to 9 points for elementary schools, 8 points for middle schools, and 7 points for high schools. (These values maintain the current percentage of schools receiving rewards at each of the three levels of schooling.)
5. A fixed target equal to 32 points (which is 5 percent of the average 1999 API score).

These options are illustrated graphically in Figure 1. All models except the fourth one involve "5 percent improvement," as required by law. They merely implement this requirement differently. The current system serves as the benchmark for comparison. The second model sets the maximum possible API score as the target, so all schools have considerable room to improve. The third requires each school to improve 5 percent of its own score, thus imposing tougher requirements on higher-scoring schools. The last two models set the same improvement target for all schools (of each type). The fourth is set at the level that would reward the same number of schools as the current system. It is far more lenient than the fifth, which is set at the API value that reflects 5 percent of the average API score for the state in 1999. Obviously, this higher target results in fewer schools receiving rewards.

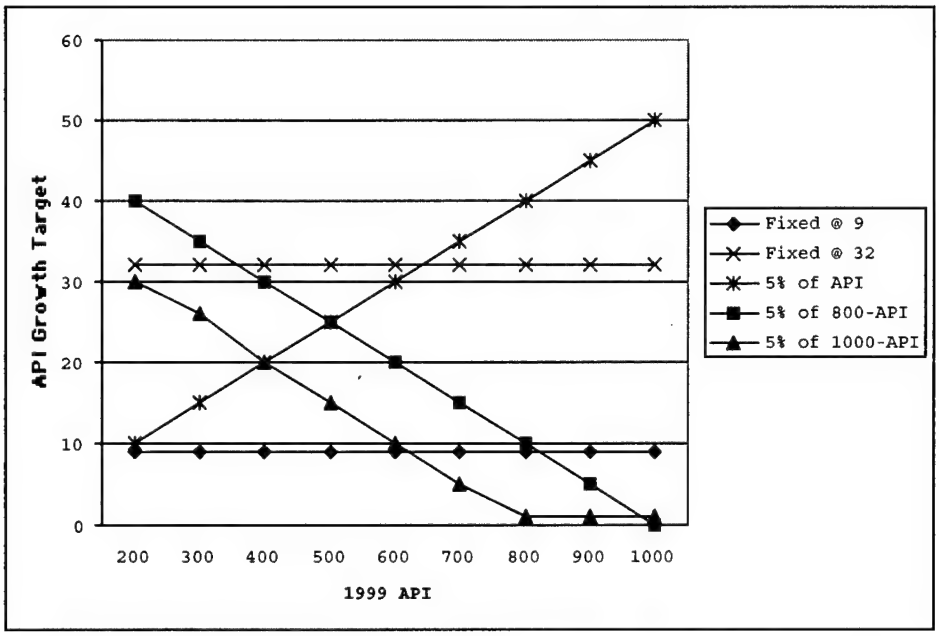

Figure 1
Alternative Growth Targets

Changing Subgroup Targets

Finally, we examine how subgroup target schemes change the allocation of school rewards. For these analyses we maintain the same rules that are currently in place regarding subgroup size and composition. We only vary the way subgroup targets are set. Similarly, we use the current rules for school-wide targets, i.e., 5 percent of the distance to 800, but we vary the subgroup targets in the following ways:

1. The current system: Significant subgroups must remain above 800 or improve by 80 percent of the school-wide target.
2. No subgroup targets at all.
3. The subgroup targets are equal to 100 percent of the school-wide target.
4. The subgroups have their own targets based on their own scores, not the school-wide score, i.e., 5 percent of the distance between the subgroup score and 800 or one point for those subgroups with 1999 API scores of 780 or higher.

Results

Changing to a Linear Metric

We expect that changing from a progressive weighting system to a linear one would benefit higher-scoring schools relative to lower-scoring ones, i.e., that the lower-scoring schools would fail to meet the target more often while the higher-scoring schools would be more likely to meet the target. Figure 2 shows the percent of schools in each quintile that would have met equivalent API and NCE improvement targets set at the 50th percentile of 1999 to 2000 actual gain. (Figure

A-1 in the Appendix shows the pattern if the criteria are set so that 70 percent of the schools meet the target.)

The difference between the percent of schools that meet the equivalent API and NCE targets increases as the base year API increases. Higher-scoring schools perform relatively better under the more linear NCE formula than under the more progressive API formula. By comparing the distance between the bars in Figure 2 it is easy to see that the API metric favors lower-scoring schools while the NCE metric favors higher-scoring schools. The only exception to this pattern is the middle quintile. We should also note that the extent of the differential impact of changing scales may be understated by this analysis because very few schools in LAUSD scored in the higher two quintiles. The pattern is the same when the target is set lower, so that 70 percent of schools succeed on both metrics, as shown in Figure A.1 in the Appendix.

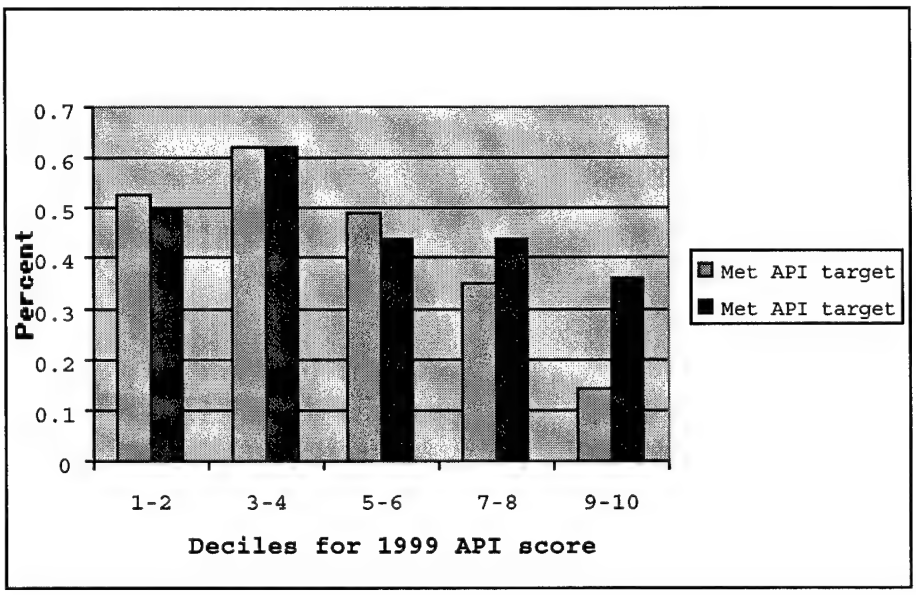

Figure 2
Percent of Schools Meeting Equivalent API and NCE Targets
(50 Percent Earning Rewards)

Changing School-wide Targets

A different pattern emerged when we looked at changes in the API targets while holding the underlying scale constant. We compared the present system ("5 percent to 800") with four other models. Figure 3 shows the percent of elementary schools in each decile (based on 1999 API scores) that received rewards under each of the five approaches. The approaches are ordered in the charts according to complexity. As the figure shows, the current system does a relatively good job of rewarding schools across the range of 1999 API scores. Over 70 percent of the schools in each decile qualified for rewards under this system, with the top two deciles achieving the highest success rates. Among middle schools and high schools the differences in favor of the top two deciles is even greater (see Figures A-2 and A-3 in the Appendix).

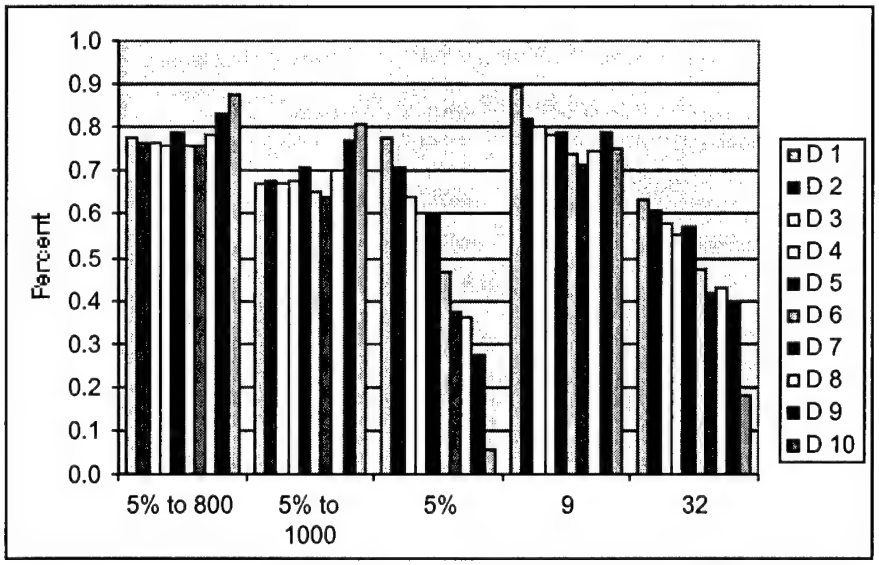

Figure 3
Percent of Elementary Schools Meeting Current and Alternative Targets by API Decile

The second set of columns reflects what would happen if the requirements were toughened for everyone by moving the benchmark score from 800 to 1,000. The percent of schools meeting the target drops in every decile, and the relative success does not change that much. Overall, changing the goal to *5 percent to 1,000* rather than *5 percent to 800* would reduce the percentage of schools meeting the criteria by 8.8, 15.5, and 14.4 percentage points, respectively, for elementary, middle, and high schools. There does not appear to be a noticeable change in the distribution of schools meeting these criteria across deciles of 1999 scores. This is easier to see in Figure 4, which shows the change in the percentages for each decile under the new schemes relative to the current system. Note in particular that the change to 5 percent to 1,000 does not affect the top two deciles more than the other eight.

The third option, a target equal to 5 percent of base score, reduces the percentage of schools meeting the criteria overall and affects high-scoring schools more than low-scoring schools. This option would decrease the number of successful schools by 30.0, 34.2, and 25.9 percentage points, respectively, for elementary, middle, and high schools. Furthermore, this target scheme would clearly hurt the high-scoring schools the most.

The next two options use a fixed value as a target for all schools. In the first case, we selected relatively low targets of 9, 8, and 7 for elementary, middle, and high schools. (These values result in the same percentage of schools meet the criteria as under the current system.) This system would reallocate the award to the lower-scoring schools (and this redistribution is actually largest for middle schools). If the fixed target were increased to 5 percent of the average API, or 32, then the higher-scoring schools would be hurt even more.

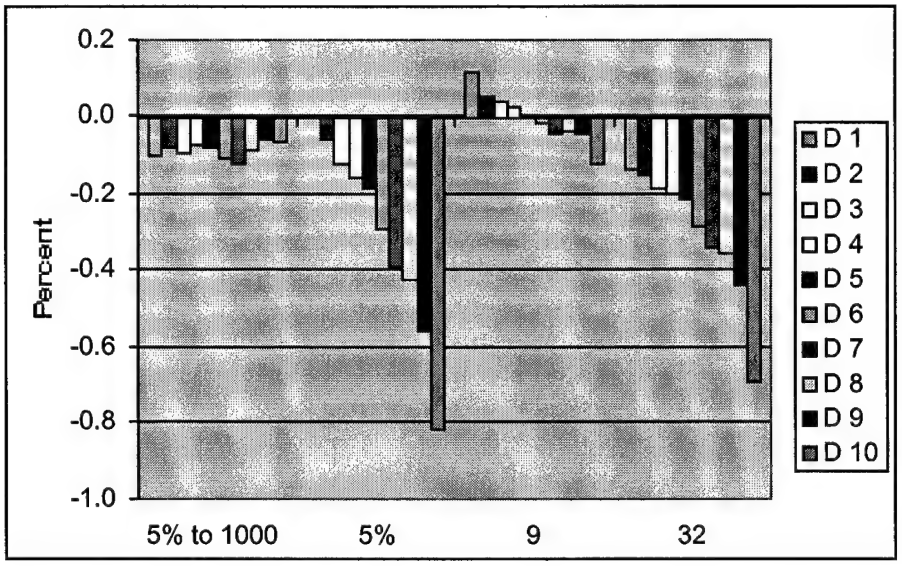

Figure 4
Difference Between Percent of Elementary Schools Meeting Current and Alternative Targets by API Decile

Figures 5 and 6 present a similar comparison of the five models based on the percent of students in the free or reduced-price lunch program. Schools are divided into quartiles for this analysis. The patterns described above are replicated when this measure of family income is used at the stratification variable. The change to a target of 1,000 affects all four groups about the same. The change to fixed targets benefits low-income schools more than high-income schools. (Similar results for middle schools and high schools are found in Figures A-4 and A-5 in the Appendix.)

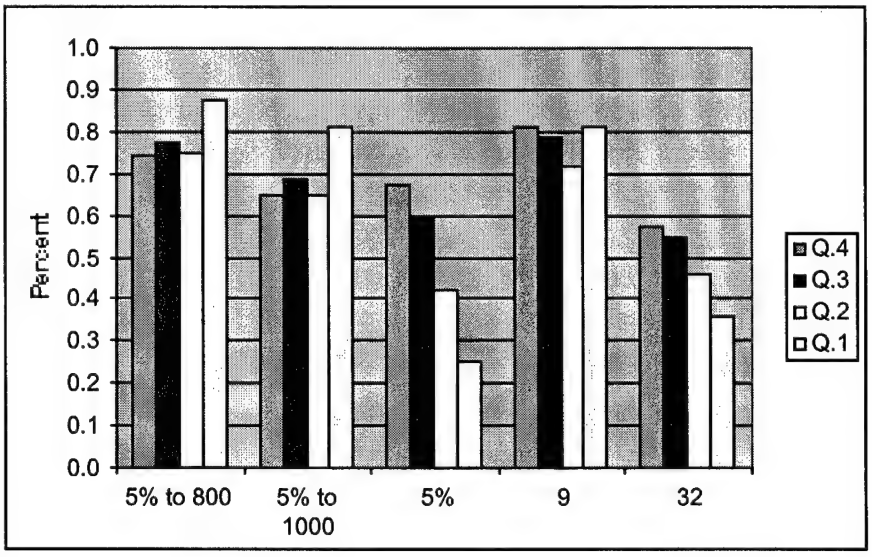

Figure 5
Percent of Elementary Schools Meeting Current and Alternative Targets by Free/Reduced-Price Lunch Quartile

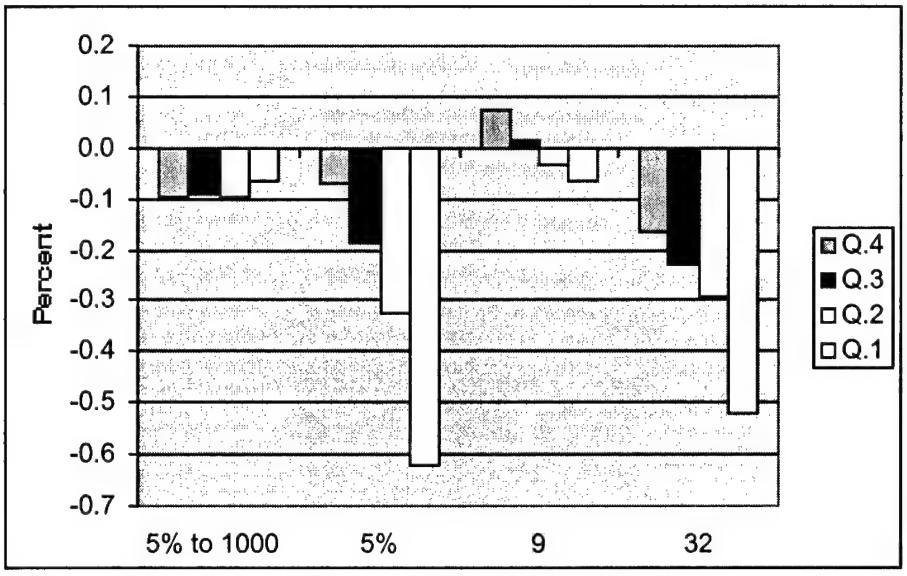

Figure 6
Difference Between Percent of Elementary Schools Meeting Current and Alternative Targets by Free/Reduced-Price Lunch Quartile

To summarize, the choice of school-wide targets has a dramatic effect on the allocation of rewards. None of the four new options considered here leads to a more equitable distribution of rewards across base year API deciles and free/reduced-price lunch quartiles than the current scheme. The use of very low fixed targets comes close in terms of the overall distribution. Moderate fixed targets reduce the rewards among the higher-scoring schools and those with fewer participants in the school lunch program more than among other schools. Having a fixed percentage improvement target also hurts the higher-scoring schools and those with fewer participants in the school lunch program.

Changing Subgroup Targets

In Figure 7 we present the percentages of elementary schools in each API decile that would qualify for reward if we changed the subgroup target rules. Figure 8 shows the corresponding change based on the percent of students in the free/reduced-price lunch program.

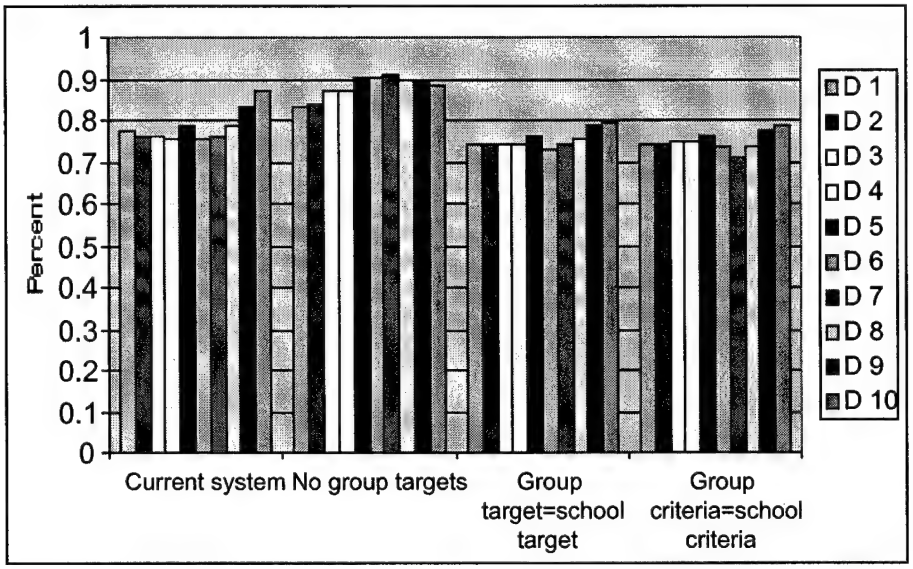

Figure 7
Percent of Elementary Schools Meeting Current and Alternative Subgroup Targets by API Decile

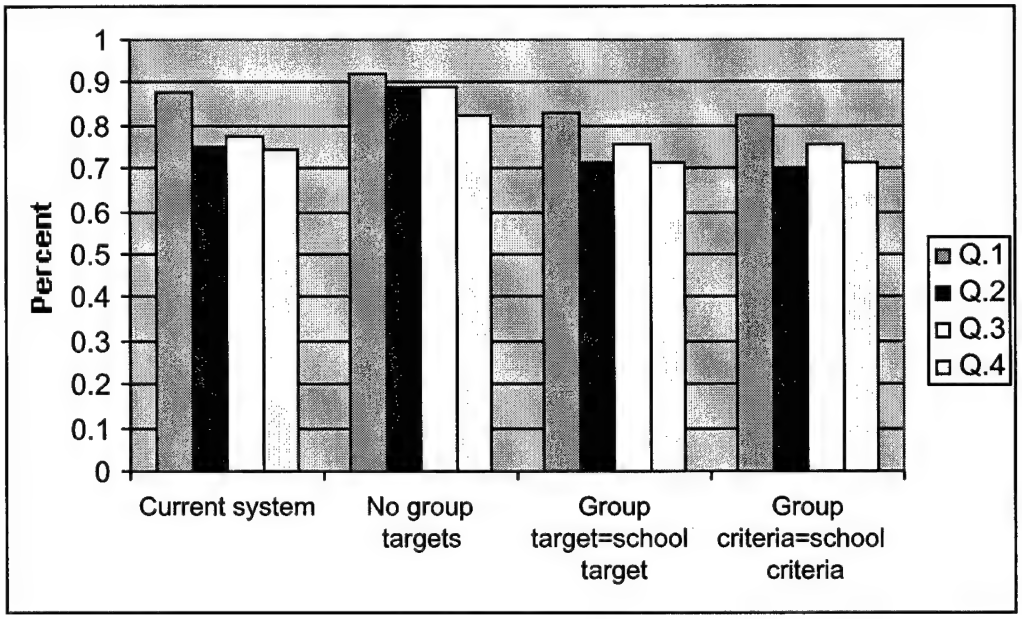

Figure 8
Percent of Elementary Schools Meeting Current and Alternative Subgroup Targets by Free/Reduced-Price Lunch Quartile

Each of the three new schemes for the subgroups would slightly reduce the disparity across the deciles and quartiles. This can be seen by the similar heights of the bars in Figure 7. The schemes would have different effects on the percentage of schools qualifying for reward. The first policy of eliminating the group targets would increase the percentage meeting the target by

9.6, 13,8, and 15.3 percentage points for elementary, middle, and high schools. For each level of schooling, it would be the schools in the "middle" of the distribution of 1999 API scores and of the distribution of percent of students on school lunch programs that would have the greatest relative increases.

The next policy of setting the subgroup targets to the same value as the school-wide targets would decrease the percentage of schools meeting the criteria by about 3 to 5 percentage points. For the most part, it is the highest-scoring schools and the schools with the lowest percentage of students in the school lunch program that would be disproportionately hurt by this program.

In the final approach, each of the significant subgroups has its individual target calculated separately using the same method that is used for the school-wide target. That is, each subgroup would need to improve by either 5 percent of the distance to 800 or one point if their score is 780 or higher). This approach would reduce the percentage of schools receiving the award by between 3 and 6 percentage points. For elementary schools, this policy would reduce the probability of meeting the criteria more for the higher-scoring schools and for the schools with fewer students in the free/reduced-price lunch program. The patterns are similar for middle schools, but not for high schools (see Figures A-6 to A-9 in the Appendix).

Discussion

The analyses illustrate that for test-based accountability systems, "the devil is in the details." The California legislature enacted what it thought was a very explicit standard of at least 5 percent improvement. However, it is possible to interpret this mandate in many different ways, which lead to many different results. This paper demonstrates that by changing the scale, the targets and the rules governing subgroups it is possible to shift rewards toward low-performing schools, share rewards relatively uniformly across all schools, or shift rewards toward high-performing schools. By describing what would have happened in 1999-2000 had different rules been applied, we hope to provide useful information for California and other states contemplating test-based accountability systems.

This analysis shows that the approach chosen by the California Department of Education appears to have achieved the state's goals reasonably well. In particular, the California model distributes rewards relatively evenly compared to the other models we examined. Only two other approaches yield more uniform rewards in certain circumstances. Using a fixed target of 9 points reduces disparities slightly with respect to free/reduced-price lunch status (but not with respect to API deciles). Setting the subgroup target equal to the school target reduces disparities slightly with respect to API deciles, but lowers the percent of schools meeting the target overall.

The perception that rewards are possible regardless of initial status is likely to keep teachers in all schools motivated to improve. It is interesting to note that the model in which we set a single fixed target for all schools yielded a distribution of rewards that was quite similar to the "5 percent of distance to target" approach adopted by the state, and it is much simpler conceptually. Yet, this may be a case where simplicity is not an advantage. The state's more circuitous formula may have some advantages from the point of view of educational improvement. By setting progressive weights which give more credit to gains in achievement at the bottom of the scale it has signaled its concern that attention be paid to those most in need. By setting greater targets for lower-achieving schools it has signaled its belief that those schools with the worst records must work the hardest.

This positive result does not constitute an unreserved endorsement of the current system, since there are many other aspects of the system we did not examine. For example, we did not explore

the effects of regulations regarding the percentage of students who must be tested, the number of years students have been enrolled in the district, etc. We had hoped to explore the impact of student mobility at the school level on rewards, but we were not able to find an appropriate dataset to study this question. We also did not look at the statistical properties of the index, which are extremely important. For example, the Technical Design Group that advises the state has examined analyses showing how the subgroup rules affect the likelihood of making false positive and false negative decisions about rewards.

Finally, it is important to continue to consider alternatives as the rewards program matures. The SAT-9 test was used for the first time in 1997-98, and there is likely to be some score inflation in the data reported here. This may decline over time, which might result in declines in the percentage of schools meeting their targets, and these declines may not be evenly distributed. On the other hand, as information about large teacher rewards becomes more widespread, it may motivate teachers to focus more energy on test-related skills leading to further increases in scores. Continuing to do analyses such as this will help the state monitor and improve its system over time.

Appendix

Table A-1
Demographic Characteristics of Schools in California by Level

	Elementary Schools		Middle Schools		High Schools	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Percent in lunch program	51.6	31.0	43.5	27.2	28.4	22.4
Percent English language learners	25.1	22.8	18.9	17.1	13.2	13.0
Percent taking the test	99.2	1.1	98.9	1.2	98.0	1.5
Percent who were new students	18.3	10.2	17.7	16.0	14.0	17.7
Percent of schools with 1 significant subgroup	15.9		12.2		13.5	
Percent of schools with 2 significant subgroups	35.6		28.6		22.3	
Percent of schools with 3 significant subgroups	39.2		39.9		36.9	
Percent of schools with 4 significant subgroups	8.7		15.9		17.0	
Percent of schools with 5 significant subgroups	0.5		3.4		7.6	
Percent of schools with 6 significant subgroups	0		0.1		2.5	

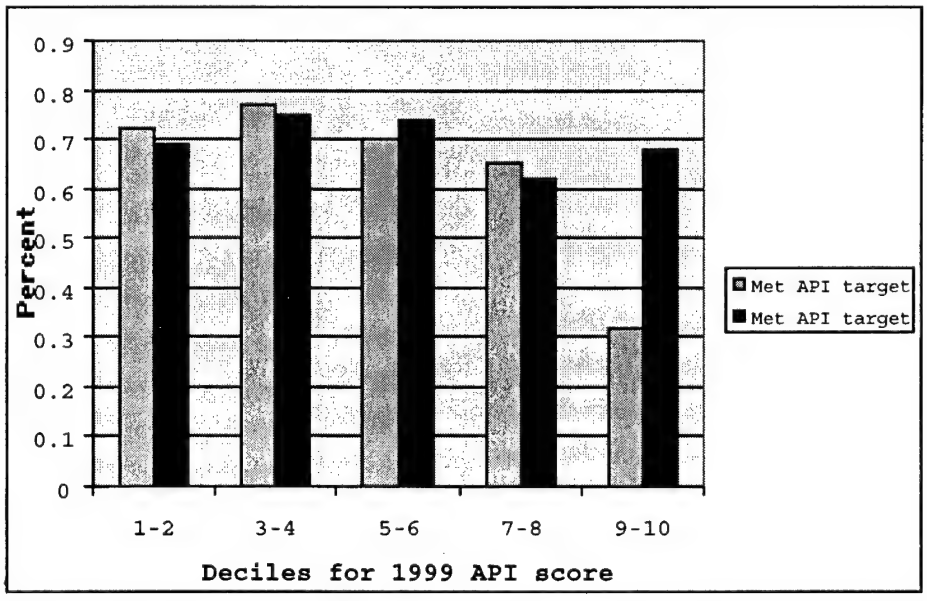

Figure A-1
Percent of Schools Meeting Equivalent API and NCE Targets
(70 Percent Earning Rewards)

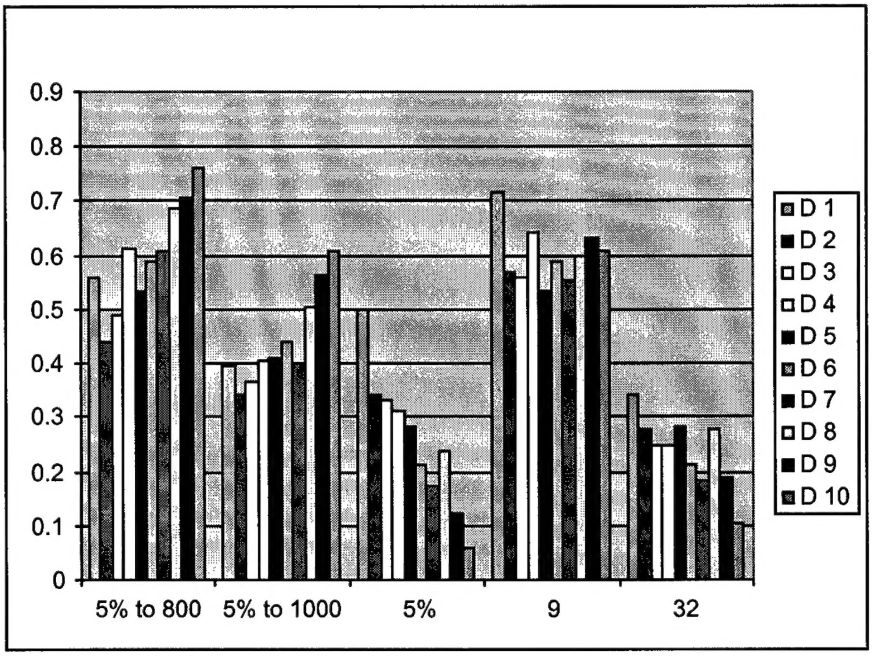

Figure A-2
Percent of Middle Schools Meeting Current and Alternative Targets by API Decile

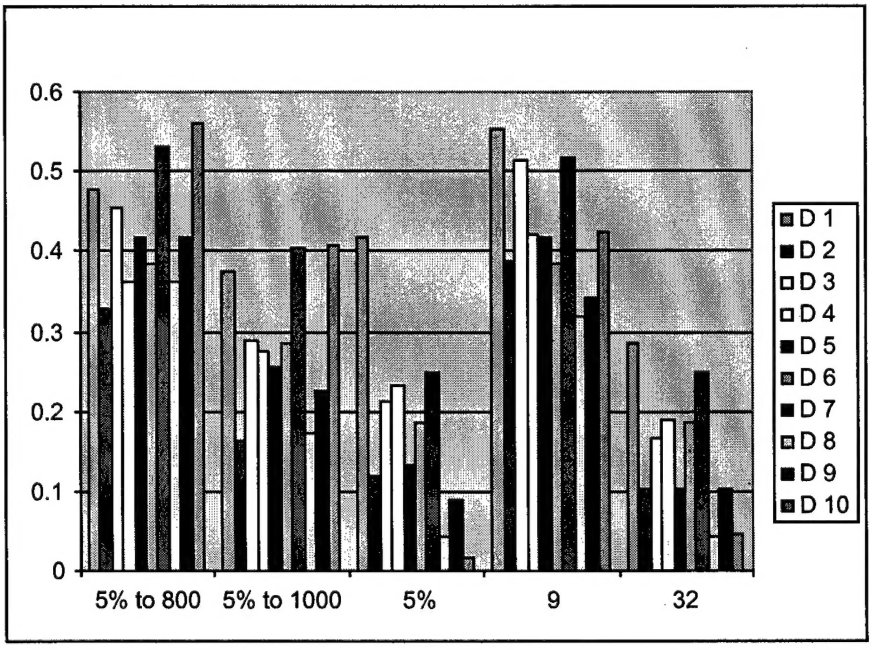

Figure A-3
Percent of High Schools Meeting Current and Alternative Targets by API Decile

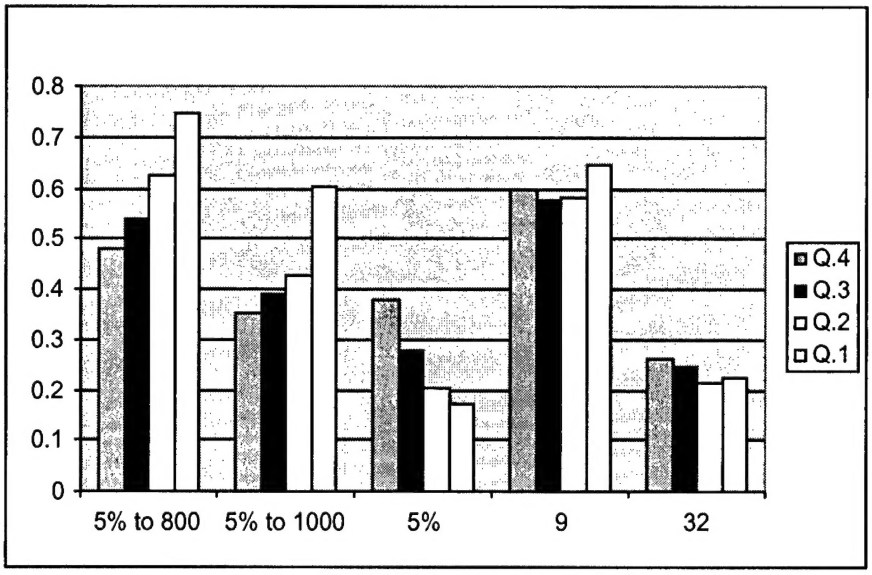

Figure A-4
Percent of Middle Schools Meeting Current and Alternative Targets by Free/Reduced-Price Lunch Quartile

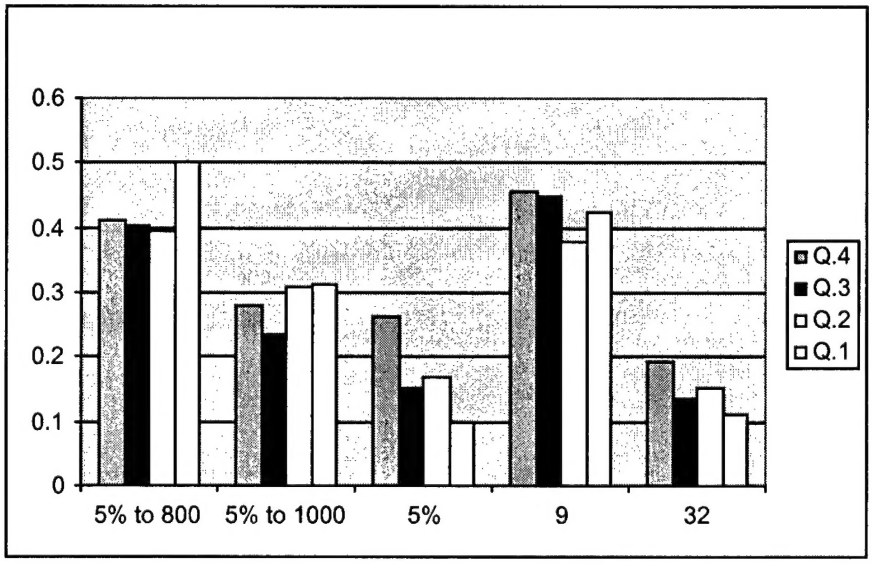

Figure A-5
Percent of High Schools Meeting Current and Alternative Targets by Free/Reduced-Price Lunch Quartile

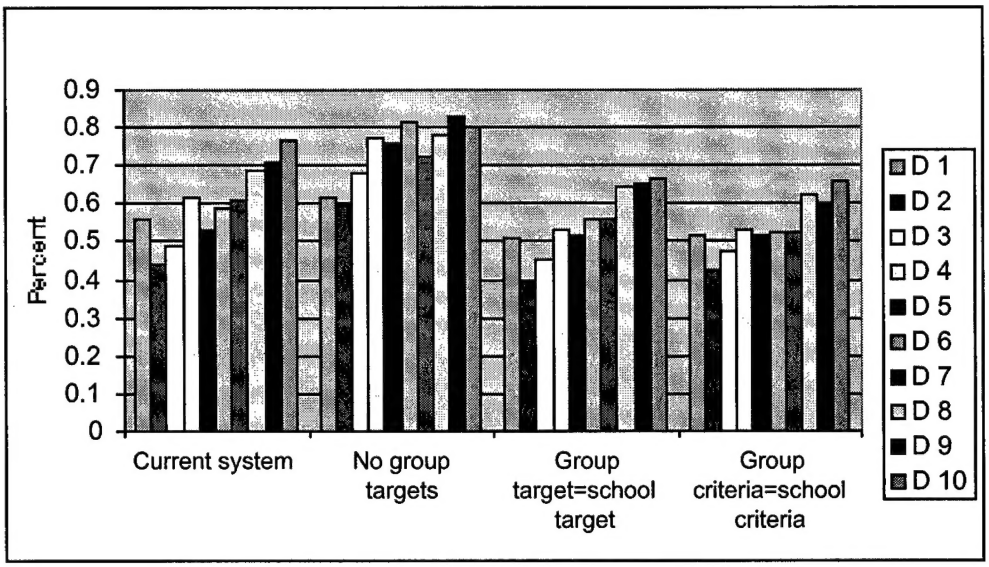

Figure A-6
Percent of Middle Schools Meeting Current and Alternative Subgroup Targets by API Decile

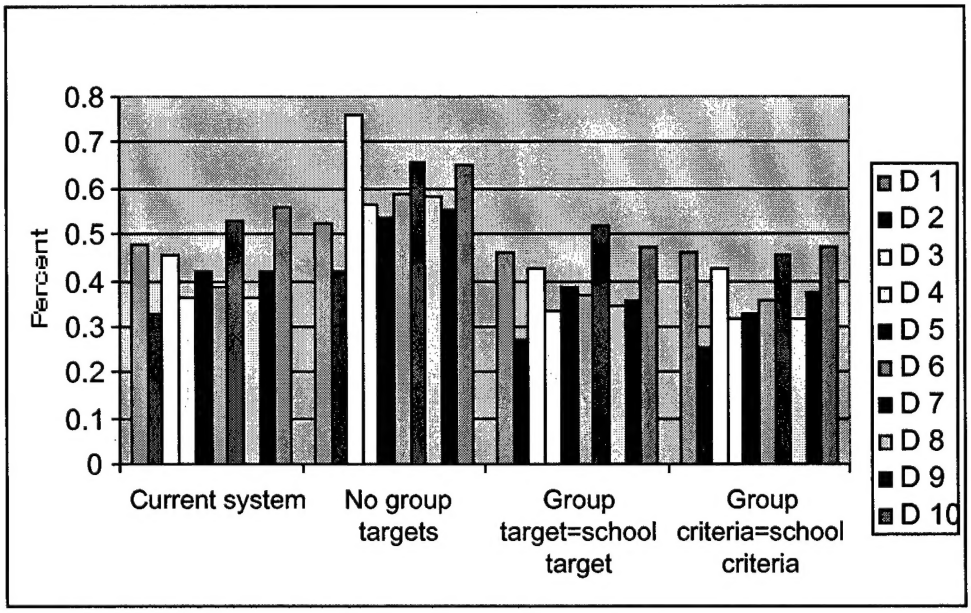

Figure A-7
Percent of High Schools Meeting Current and Alternative Subgroup Targets by API Decile

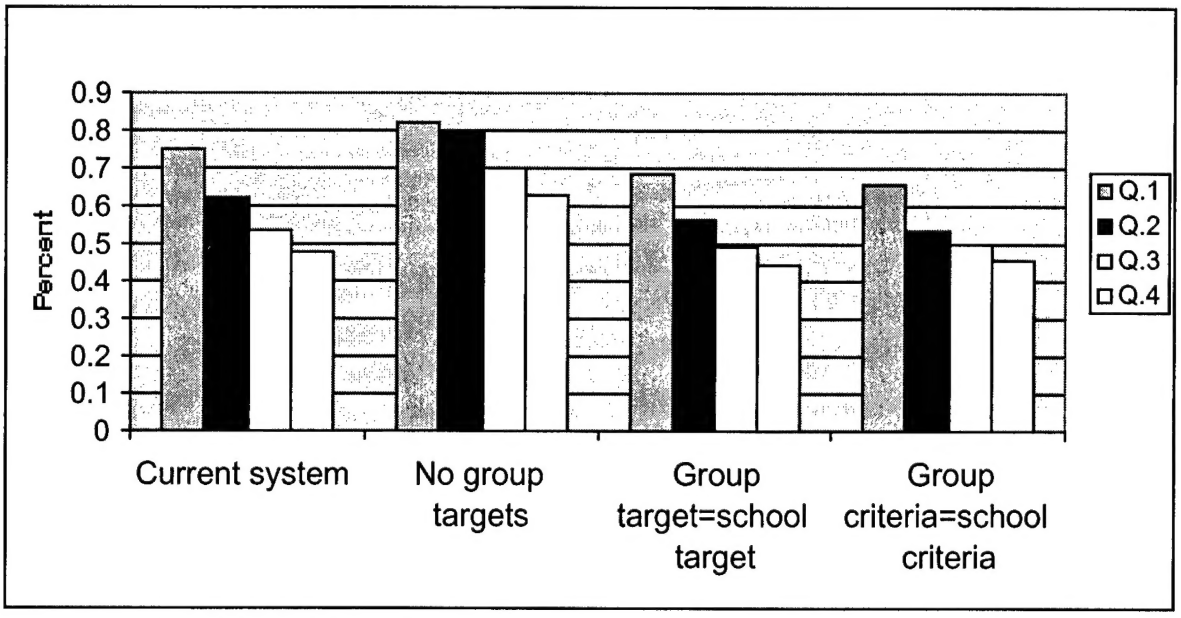

Figure A-8
Percent of Middle Schools Meeting Current and Alternative Subgroup Targets by Free/Reduced-Price Lunch Quartile

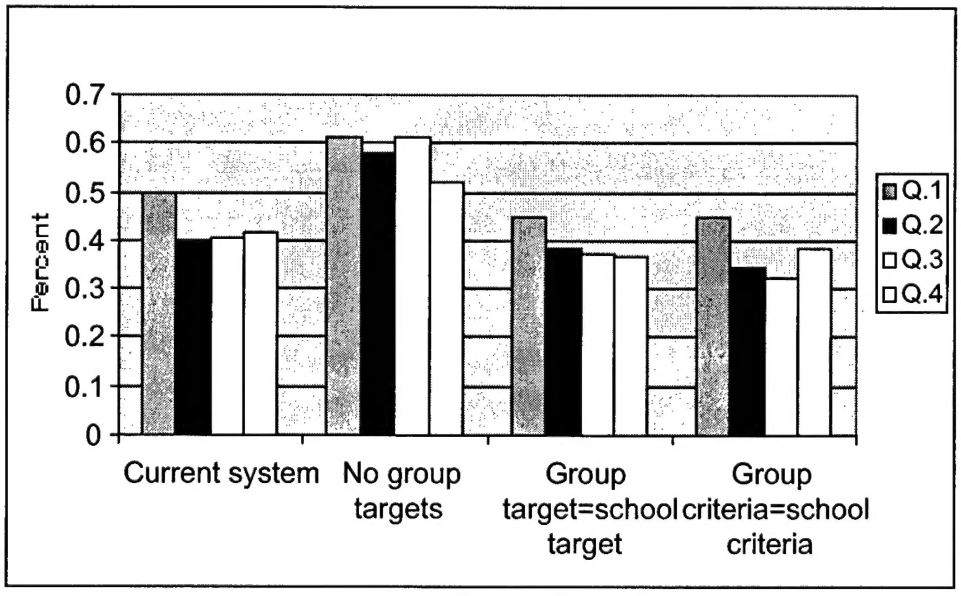

Figure A-9
Percent of High Schools Meeting Current and Alternative Subgroup Targets by Free/Reduced-Price Lunch Quartile